Version 8.0
September 18, 2015

Evaluation Plan for the NIST Open Evaluation of Speech Activity Detection (OpenSAD15)

1. Introduction

The NIST open evaluation of Speech Activity Detection (OpenSAD) is intended to provide SAD system developers an independent evaluation of performance on a variety of audio data. The intention of this evaluation is to advance technology that can be used to pick out regions of speech in an audio file for a human user to examine and also for downstream automatic processing by technologies such as speech recognition, speaker identification, language identification, or machine translation. We want the evaluation to also be useful to researchers who are trying to model the characteristics of speech rather than modeling noise (noise that could, for example, be subtracted from the signal leaving just regions of speech or of other "source" sounds) — we are not, however, interested in restricting the approaches that can be taken to SAD.

The underlying model of use is that a human wants to find areas (that are possibly rare) containing speech within large volumes of audio data.

The NIST OpenSAD evaluation has ties to the DARPA Robust Automatic Transcription of Speech (RATS) program, in which the evaluations were only open to the RATS performer teams. The RATS program was designed to advance the current state- of-the-art in identifying speech activity regions in signals from distorted, degraded, weak, and/or noisy communication channels[1]. Most of the data for the OpenSAD evaluation will match that description.

We anticipate that at least some system developers may build systems that can do unsupervised learning/adaptation/modeling. In order to support such research, the OpenSAD evaluation will report scores that allow participants to compute (for such systems) the relative gain of adapted systems (having done adaptation on the evaluation dataset) compared to a baseline of the un-adapted system trained on the RATS training dataset. Such adaptation will not be a required element of the evaluation. Developers who wish such scores will of course submit separate outputs for those two conditions.

One goal is to enable the research community to understand the accomplishments of the participating systems (what worked, what did not). Accordingly, for the workshop at the end of OpenSAD all participants must provide a system description that explains the approaches taken. In their workshop presentation, teams can tie their results to that.

[1] See the following for additional information on the DARPA RATS program.
http://www.darpa.mil/Our_Work/I2O/Programs/Robust_Automatic_Transcription_of_Speech_(RATS).aspx
https://www.fbo.gov/?id=5e55f8e0a990ebde2c197c3c6387b9f7

For systems that perform adaptation, the system description should include whether the adaptation was over the entire dataset or was adaptation to the local waveform (noise and/or speech in the local region). The evaluation framework stated in this evaluation plan is intended to enable participants to do research that identifies the adaptation strategies that are most successful and characterize the tradeoff between adaptation effort and performance gains. Systems doing adaptation over the entire dataset should not adapt on any files that are from the "XMT" channel (the clean data that was input to the transmitter) nor should they adapt over the low-speech-density-data (LSDD) that is described in the next section.

2. Data

We will use sequestered RATS evaluation set (from LDC), as a baseline. Some of that data includes noises, a variety of transmitter/receiver radio-link channels, and data in which the regions of speech are a small part of the overall time. The speech data in the OpenSAD evaluation will have originated as telephone speech (landline or cell) over public telephone networks (for languages other than English, sometimes one party was in the U.S. and one party was in some other country).

At the end of the evaluation dataset, there will be a separate low-speech-density data (LSDD) evaluation dataset that mostly consists of various sorts of noises, transmitted over channels used in other evaluation data; and for the LSDD dataset only the non-speech regions (thus false-alarms vs. true-negatives) will be scored.[2] Systems should not do supervised adaptation to the LSDD data and should not carry-over unsupervised adaptation from one LSDD file to the next (process each LSDD file as if it were the first).

The *training datasets* will consist of speech transmitted-and-received via several different types of radio channels, each channel having individual noise, bandwidth, and distortion characteristics. *Devtest datasets* from these same channels will also be provided, and participants can use them to test or tune their systems. These *training* and *devtest* datasets will be provided early in the program. There will be no training or devtest datasets for the LSDD dataset.

Because we want to support participants who wish to assess their ability to model characteristics of speech rather than just model the (known) characteristics of noise or speech distortion, parts of the *evaluation dataset* for OpenSAD may include noise (possibly varying noise) characteristics and speech-distortion (possibly varying distortion) characteristics that substantially differ from the *training* and *devtest* datasets[3].

2.1 Statistical properties needed

To ensure statistical confidence in reported results, the evaluation dataset needs to be large enough to have a substantial number of errors at the anticipated error rates. At error rates of 3% miss and 1% false alarm, if we were to try for 50 miss errors and 50 false alarm errors, this would imply 1,667 speech intervals and 5,000 non-speech. This is, of course, a discussion of statistical design, *not* a statement about the actual evaluation datasets.

[2] Regrettably, we lack the resources to accurately annotate the speech regions on the **received** signal on the various channels for this particular dataset. Because we can annotate the regions of speech on the **transmitted** signals, we can omit them from the evaluation.

[3] So far, we do not have such data, suitably annotated; no promise is being made.

Non-speech intervals may be regions annotated as substantially different from each other, rather than just regions separated by intervals of speech. For a portion of the Evaluation dataset, the occurrence of speech may be rare[4]. Most of the data will include various channel distortions.

2.2 Annotation

NIST will undertake to have a careful annotation of all Training, DevTest, and Evaluation datasets. There is no assumption in this evaluation that the speech will be English or even in a language recognizable by the participants; there will be a range of human languages. In regions where the annotator is `uncertain` whether there is speech, the annotation may so indicate[5], and the scoring will omit such regions from scoring. Similarly, the annotation may mark `NT` for regions where there is a gap in the transmission, and such regions will be scored as non-speech (in effect, as silence).

2.3 Datasets

There will be three types of datasets: a *Training* dataset (the LDC-released SAD training datasets), substantial *DevTest* dataset(s), and *Evaluation* datasets.

3. Evaluation setup

The evaluation datasets will be provided to the participants for the evaluation, and participants will run the evaluation datasets through their systems to generate output. NIST anticipates that licensing agreements for the evaluation datasets will require participants to delete the evaluation datasets after the evaluation (and return the media on which the datasets are distributed). Participants will deliver their system outputs to NIST, and NIST will score them.

4. Performance measures

SAD error rates are estimated from the amount of time that is misclassified in a system segmentation of test audio files.

For OpenSAD, missing (failing to detect) actual speech is considered a more serious problem than having a region of speech identified as beginning a little before it actually begins and/or as ending a little after it actually ends. Accordingly, as the official metric we will allow systems a two-second collar at the beginning and end of each speech region, within which we will not score false-alarm errors (notice that this implies that a region of non-speech lasting less than four seconds will not be scored because it will be subsumed by the [merged] collars). The scoring software will also provide scores with shorter collars (collar lengths of 1 second, 0.5 seconds, 0.25 seconds, and no-collars[6]), as additional feedback to participants, and the size of a non-speech region where collars will merge will (of course) vary accordingly.

Although the collar sizes will merge as described in the preceding paragraph, the scoring script actually does a tweak to that[7], as follows. Assuming a two-second collar, if a non-speech region lasts just barely over four seconds then the scored non-speech region between the two collars would (as a result) be very short. Similarly for other collar sizes. The tweak is that if such a segment of non-speech between collars will not last at least a tenth of a second (0.1 sec) then the

[4] Some files may contain ***no*** speech at all.

[5] This is hypothetical. No such segments occur in the annotation for the data currently on hand.

[6] This means "no collars" rather than "collars of length zero."

[7] This tweak does not apply for the no-collar scores.

collars involved will expand so that they will still merge (for example, no resulting non-speech segment with a duration of just 0.099 seconds). Similarly for a region of non-speech before a collar at the beginning of the file or a region of non-speech after a collar at the end of the file the resulting non-speech segment must last at least a tenth of a second or else the collar will expand. In all other circumstances the collars will be exactly the nominal length.

Figure (1) illustrates the relationship between human annotation, the scored regions that result from application of the collars, a possible system output, and the resulting time intervals scored as:

- true negative TN (correctly identified regions of non-speech),
- true positive TP (correctly identified regions of speech),
- miss, or false negative FN, and
- false alarm or false positive FP time.

The scoring collars also compensate for ambiguities in noisy channel annotation. Non-speech collars of two seconds in length, shown above the annotation, define regions that will not be scored. As can be seen, collars are applied to the annotations to determine the parts of the speech and non-speech that are scored.

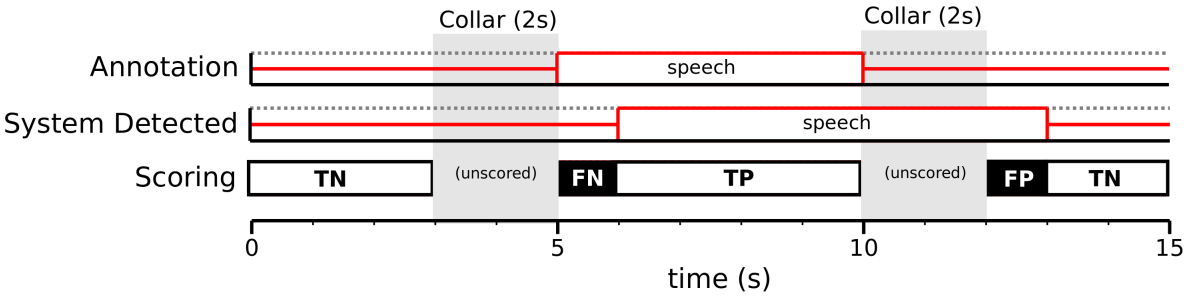

Figure 1: After collar application, systems are not scored on false alarms within two seconds from speech boundaries.

In theory, each segmentation (system output over the dataset) represents a single operating point in a detection error trade-off space. Nevertheless, the OpenSAD evaluation will ***no***t generate DET curves or ROC curves.

For each system output for each input file, two error rates will be calculated. Scored regions define the denominators in the miss and false alarm rate equations shown below. The presentation of results will discuss performance on both.

- Miss rate(or P_{Miss}) $= \frac{total\ FN\ time}{total\ speech\ time}$

- False alarm rate (or P_{FA}) $= \frac{total\ FP\ time}{total\ scored\ nonspeech\ time}$

As clarification of the above formulas, note that the *total speech time* is determined by the reference annotation, and it will equal the *total TP time + total FN time*. Test data will vary in how much speech is present in each sample (that is, some files will have more speech, some less or even none). As can be seen in Figure 1, the collars will not affect the *total speech time.*

Similarly, the *total scored nonspeech time* is *total FP time + total TN time*. But the *total scored nonspeech time* does not include the unscored collar time, and therefore **will differ for the alternative scorings using shorter collars**.

Systems under test will be evaluated on all test samples.

With P_{Miss} and P_{FA} as defined above, system developers should minimize the following *Detection Cost Function* metric:

$$DCF(\Theta) = 0.75 \times P_{Miss}(\Theta) + 0.25 \times P_{FA}(\Theta).$$

In the *DCF* formula, Θ is the operating point defined by the SAD system's internal weights and thresholds, 0.75 is the *Cost* of a *Miss*, and 0.25 is the *Cost* of a *FalseAlarm*.

At the request of participants, DCF (with the 0.75 and 0.25 costs stated above) is the official metric for the OpenSAD evaluation.

5. Data formats

The audio data format(s) to be processed will be .flac format, 16 bit, 16k/sec., audio files.

The files to be processed will be specified by an XML file that defines the test. Figure 2 is an example of that Test Definition file format. RATS participants should note that this file begins with a `TestSet` tag rather than `RATSTestSet`.

In the SAMPLE element: the `id` attribute's value ties the Test Definition to the system output, and the `file` attribute is a filename in that directory, usually with a directory path (relative to the current directory).

```
<TestSet id="OpenSAD" audio="/path/to/audio/root" task="SAD">
   <TEST id="SADTestDataset1">
      <SAMPLE id="SAD_sampleFile1" file="set1/G/file1.wav" />
      <SAMPLE id="SAD_sampleFile2" file="set1/G/file2.wav" />
      ...
   </TEST>
</TestSet>
```

Figure 2. Example of the Test Definition file format, which is XML

System outputs will be a tab-separated ASCII text file with nine columns. Figure 3 defines that file format. The "answer key" annotation, that outputs will be scored against, shown in Figure 4, is similar to this.[8]

[8] RATS participants should note that the Type (field 8 in Figure 3) did not occur in RATS. We want to allow systems to state confidence for regions of non-speech, not just for speech.

Column	Description
1: Test	Test Definition File name (name of the file whose content is illustrated in Figure 2)
2: TestSet ID	contents of the `id` attribute of the `TestSet` tag (see Figure 2)
3: Test ID	contents of the `id` attribute of the `TEST` tag (see Figure 2)
4: Task	`SAD` <== a literal text string, without quotation marks
5: Sample ID	contents of the `id` attribute of the `SAMPLE` tag
6: Interval start	an offset, in seconds, from the start of the audio file for the **start** of a speech/non-speech interval
7: Interval end	an offset, in seconds, from the start of the audio file for the **end** of a speech/non-speech interval
8: Type	In system output: "speech" or "non-speech" (with no quotation marks). In the reference: S, NS, or NT (for Speech, Non-Speech, and NoTransmission).
9: Confidence (*optional*)	A value in the range 0.0 through 1.0, with higher values indicating greater confidence about the presence/absence of speech

Figure 3. Format of the SAD system outputs

Column
1: Audio filename
2: Channel ID
3: Interval start time
4: Interval end time
5: Type (S, NS, or NT)
6: SAD provenance (generally manual, meaning manually annotated)
7–12: Not relevant to SAD

Figure 4. Format of the SAD annotation (answer key) files

In the SAD annotation, Field 5 (Type) may have the values `S` (for speech), `NS` (for non-speech), or `NT` for a gap in the transmission. It is also (hypothetically) possible to have the value "`uncertain`" (as explained in section 2.2). The correct annotation file is the one that matches on both Field 1 (the audio filename) and Field 2 (the channel ID). Field 6 (provenance) relates to the source of the information, and is not relevant to scoring.

Note:

The NIST website for OpenSAD includes a set of mock files, including an `example_testDefFile.xml` file, a set of mock system output files, a corresponding set of mock answer key files, and beta-version scoring software that will process those files. Many possible questions about file formats and naming may be resolved by examining those examples.

6. Planned Schedule

Sept 3: Publish registration form, mock data, and beta-version scoring code

Sept 7: Release Data-licensing agreements, from LDC

Sept 21: Release Training and DevTest datasets, to registered participants

Oct 1: Participant registration deadline

Nov 2-13: Evaluation period
(Two weeks to accommodate multi-site teams)

Nov 26: Release preliminary results (automated, non-refereed)

Dec 18: One-day workshop (date changed from Dec. 17, which conflicts with ASRU)

Dec 31: Latest date for NIST to publish final results
(with any refereeing or updates triggered by the workshop)

7. Change Notes (version history of the Evaluation Plan)

Version 7.0 was the initial wide public release

Version 7.1 mentioned that the workshop would probably change to Dec. 18

Version 8.0 changes the official metric to DCF (see pg. 5),
clarifies the LSDD adaptation rules (see pg. 2),
and makes definite that the workshop date is Friday Dec. 18